Report Number: GA22F042

Final Report: June 1, 1999

Electronic Digital Imaging Standards for Archiving Records

By

Susanne H. MacTavish and Michael R. Pickard

Lockheed Martin Technology Services

Information Support Services

Falls Church, Virginia

LOCKHEED MARTIN

Contract: GS-35F-4863G

Lockheed Martin Reference Number: GA22F042

Approved for public release; distribution is unlimited

(This space intentionally left blank.)

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	01 June 1999	Final

4. TITLE AND SUBTITLE
Electronic Digital Imaging Standards for Archiving Records

5. FUNDING NUMBERS
Contract: GS-35F-4863G/GA22

6. AUTHOR(S)
Susanne H. MacTavish and Michael R. Pickard

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lockheed Martin
5203 Leesburg Pike - Suite 1500
Falls Church, VA 22041

8. PERFORMING ORGANIZATION REPORT NUMBER
GA22F042

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
OSD/C3I

Room 7012, Crystal Mall 3
Arlington, VA 22202

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
DoDD5230.24 Distribution Statement A: Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

To improve access, distribution, and interoperability, Federal agencies are converting large numbers of documents from paper to electronic digital images. Increased accessibility to the most current data drives the move away from paper records whenever possible. Among Federal agencies there is increasing interest in receiving National Archives and Records Administration (NARA) guidance identifying acceptable Digital image formats for long term preservation.

In June 1998, the Office of the Assistant Secretary of Defense Command, Control, Communications, and Intelligence (OASD/C3I) awarded a Task Order (Imaging Standard Policy Support, GS-35F-4863G/GA22) to Lockheed Martin to continue its study of digital imaging standards for archiving records. Under this Task Order the Lockheed Martin team has gathered information from the literature, interviews, and consensus gathering sessions, with a focus on three specific categories of documents that have traditionally been transferred to NARA for long-term preservation: personnel records; manuals, standards, directive type material; and documents scheduled for declassification or redact items. This report documents the current status of the three study focus areas, and provides information about digital image format options, along with associated cost and migration strategies.

14. SUBJECT TERMS
Digital Imaging Standards
DoD-NARA Scanned Images Standards Conference

15. NUMBER OF PAGES
96 + x

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	SAR

Standard Form 298 (Rev. 2-89) (EG)
Prescribed by ANSI Std. 239.18

(This space intentionally left blank.)

Executive Summary

In 1996, the Department of Defense (DoD) initiated a study of digital image format standards. The goal was to identify and evaluate alternative electronic digital image standards for the storage and retrieval of DoD digital image records. The report of that study, Electronic Imaging Standards for Archiving Records, was issued on May 31, 1997. The report recommended that DoD pursue a strategy of adopting image standards that are embodied in commercial off-the-shelf products.

Recognizing that a significant volume of DoD records, that had been traditionally transferred to the National Archives and Records Administration (NARA) as paper records, are now being created in or converted to digital image formats, DoD asked NARA to participate in its continuing imaging standards study. Thus ensuring that NARA's long-term preservation and access needs, as well as DoD's operational record requirements, are addressed.

In June 1998, the Office of the Assistant Secretary of Defense Command, Control, Communications and Intelligence (OASD/C3I) awarded a Task Order (Imaging Standard Policy Support, GS-35F-4863G/GA22) to Lockheed Martin to continue the study, and initiate actions required for implementation of selected recommendations made in the 1997 report. Under this Task Order the Lockheed Martin team has gathered information from the literature, interviews, and consensus gathering sessions, with a focus on three specific categories of documents that have traditionally been transferred to NARA for long-term preservation:

- personnel records,
- manuals, standards, directive type material, and
- documents scheduled for declassification or redact items.

It is clear that Electronic Records have become a very HOT topic.

- The use of computers is changing the way government documents are created, accessed and managed. Electronic records, the Internet and E-mail have become an increasingly large part of the everyday work environment. To improve access, distribution, and interoperability, Federal agencies are converting large numbers of documents from paper to electronic digital images. Increased accessibility to the most current data drives the move away from paper records whenever possible. Among these Federal agencies there is increasing interest in receiving National Archives and Records Administration (NARA) guidance identifying acceptable digital image formats for long term preservation.
- Long-term preservation of digitally imaged records has become problematic for Federal records requiring permanent retention. While the advantages of digitally imaged documents are tremendous, due to the relatively short life cycle of digital image technology (both hardware and software), it is commonly accepted that all formats used today will eventually become obsolete.
- Computer tapes and disks deteriorate, and the hardware and software systems on which they can be read become obsolete. For an electronic record long term preservation requires that as the technology changes that the record be migrated from one format to another and then verified to ensure no loss of data. Limiting the number of image formats to monitor for technology change becomes an essential part of long-term preservation strategy. Identification of appropriate and relatively stable formats is key to success.
- While there are currently no digital image formats that are acceptable for long-term preservation, the goal is to identify formats that are likely to live longer than others in guidelines as approved data preservation formats. By selecting such standards, agencies will be able to reduce the frequency of data reformatting required to migrate data through different standards and technology and thus to minimize the cost of digital image data preservation.

Study Conclusions:

1. Access and response to Freedom of Information Act (FOIA) requests are facilitated through electronic digitalization of records.

2. No de jure standard for digital images has reached the desired maturity level for archival purposes.
3. The hardware and software technology required for the use of digital images changes rapidly.
4. Migration costs associated with archiving of digital images of textual material are unknown.
5. The anticipated high cost associated with long-term maintenance of digital image records mandates careful screening and selection of only the most valuable digital imaged records to be accessioned into the National Archives.
6. Metadata standards have been developed, but no one standard has emerged as the most universally accepted standard for electronic image records.
7. Tag Image File Format (TIFF) and Portable Document Format (PDF), both de facto standards, are the most widely used formats for text records.
8. Joint Photographic Experts Group (JPEG), a de jure standard, is the most widely used compression standard.
9. The use of proprietary standards for producing and storing images is much more common than the use of official standards.
10. Organizations will continue to use the proprietary imaging formats due to the costs involved.
11. The key roadblock to a successful digital imaging program is the high costs associated with the program and the lack of management understanding to the need for appropriate funding in the area.
12. The lack of a format standard is no longer seen as a major issue.
13. A united government voice was needed, with strong NARA leadership and a means of sharing data.

The following phased implementation approach received general acceptance at the DoD-NARA Scanned Images Standards Conference:

1. Manage the process (records management, management and policy).
2. Study, plan, gather information through cost/benefit analysis of entire life-cycle (especially document preparation, searching, and migration).
3. Pick an interim standard during step 2, which will be accepted and supported by DoD and NARA – this will enable the cost-benefit analysis to be conducted.
4. Practice migration and preservation while documents are in active use.

Study Recommendations:

1. Image electronic digital material in the most stable formats available preferably using the latest version, but no more than two generations prior to the latest. (e.g., for TIFF image produced in January 1999 that would be TIFF version 6, 5 or 4).
   a. Image personnel records using TIFF for archiving, TIFF or PDF formats for access. Convert all current imaged records to one standardized TIFF format.
   b. Image declassified records using TIFF for archiving, TIFF 6 or PDF formats for access. Convert declassified versions of historically significant records to paper, microfilm or ASCII formats.
   c. Image manuals, standards, directive type material using TIFF, ASCII and ASCII SGML or XML tagged files for archiving. Use PDF, HTML or XML formats for dissemination.
2. Plan and budget for migration of digital images every 3-5 years with of cost equivalent to 50 – 100% of the costs associated with original imaging project.
3. Convert documents that require long-term preservation from application format to an image format.
4. Develop standard set of access metadata of textual digital images using DoD 5015.2-STD, EAD, and Dublin Core as minimum set.

5. Work with Association for Information and Image Management (AIIM) and American National Standards Institute (ANSI) to standardize TIFF header data.
6. Work with NARA to:
   a. Establish criteria for selection of digital images for accessioning in the National Archives.
   b. Accession digital images that have been imaged in the most stable format available and those that meet the selection criteria.
   c. Establish guidelines describing metadata that must accompany digital image when submitted for archival accessioning.
   d. Study and evaluate migration strategies applied to digital data archives to application in the maintenance of textual digital images.
   e. Study and evaluate formats designed for non-textual material, e.g. photography, aerial imagery, x-rays, radar, for compatibility with textual digital image formats in the archive environment.
   f. Study and evaluate de jure interchange formats for long-term archive acceptance and application in the field.

(This space intentionally left blank.)

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

(This space intentionally left blank.)

1 Summary

To improve access, distribution, and interoperability, Federal agencies are converting large numbers of documents from paper to electronic digital images. Increased accessibility to the most current data drives the move away from paper records whenever possible. Among Federal agencies there is increasing interest in receiving National Archives and Records Administration (NARA) guidance identifying acceptable digital image formats for long term preservation.

In June 1998, the Office of the Assistant Secretary of Defense Command, Control, Communications and Intelligence (OASD/C3I) awarded a Task Order (Imaging Standard Policy Support, GS-35F-4863G/GA22) to Lockheed Martin to continue its study of digital imaging standards for archiving records. Under this Task Order the Lockheed Martin team has gathered information from the literature, interviews, and consensus gathering sessions, with a focus on three specific categories of documents that have traditionally been transferred to NARA for long-term preservation: personnel records; manuals, standards, directive type material; and documents scheduled for declassification or redact items. This report documents the current status of the three study focus areas, and provides information about digital image format options, along with associated cost and migration strategies.

2 Introduction

The use of computers is changing the way government documents are created, accessed and managed. Electronic records, the Internet and E-mail have become an increasingly large part of the everyday work environment. To improve access, distribution, and interoperability, Federal agencies are converting large numbers of documents from paper to electronic digital images. Increased accessibility to the most current data drives the move away from paper records whenever possible. Among Federal agencies there is increasing interest in receiving National Archives and Records Administration (NARA) guidance identifying acceptable Digital image formats for long term preservation.

Title 44 of the United States Code (USC) and Title 36 Code of Federal Regulations (CFR) clearly identify the roles and responsibilities of federal agencies and the National Archives and Records Administration in the preservation of records of national historical interest.

Title 44 USC provides the NARA authority. It assigns the Archivist of the United States the responsibility to provide guidance and assistance to Federal officials on the management and disposition of records, to store records in centers from which agencies can retrieve them, and to take into archival facilities and Presidential libraries, for public use, records that are, in the language of Section 2107, "determined by the Archivist of the United States to have sufficient historical or other value to warrant their continued preservation by the United States Government."

As defined in Section 3301, these records are -- all books, papers, maps, photographs, machine readable materials, or other documentary materials, regardless of physical form or characteristics, made or received by an agency of the United States Government under Federal law or in connection with the transaction of public business and preserved or appropriate for preservation by that agency or its legitimate successor as evidence of the organization, functions, policies, decisions, procedures, operations, or other activities of the Government or because of the informational value of data in them.

Title 36 Code of Federal Regulation (CFR) section 1234 sets the rules for agencies to follow regarding Electronic Records. It states that agencies must address electronic record management and that NARA should be a player in deciding how they manage their electronic records. Agencies are required to select appropriate media and systems for storing the agency's electronic records through out their life. It further states that while an agency does not need to store records in media and formats specified in 36 CFR 1228.188, it must be willing and ready to migrate the records to the currently required transfer media and formats for all permanently valuable electronic records. 36CFR1228.188 d Formats (2) Textual documents states, "Electronic textual documents shall be transferred as plain ASCII files; however, such files may contain Standard Generalized Markup Language (SGML) tags."

For Federal records requiring permanent retention, long-term preservation of digitally imaged records has become problematic. While the advantages of digitally imaged documents are tremendous, due to the relatively short life

cycle of digital image technology (both hardware and software), it is commonly accepted that all formats used today will eventually become obsolete.

Computer tapes and disks deteriorate, and the hardware and software systems on which they can be read become obsolete. Long term preservation requires that as the technology changes an electronic record must be migrated form one format to another and then verified to ensure no loss of data. Limiting the number of image formats to monitor for technology change becomes an essential part of long-term preservation strategy. Identification of appropriate and relatively stable formats is key to success.

While there are currently no digital image formats that are acceptable for long-term preservation, the goal is to identify formats that are likely to live longer than others in guidelines as approved data preservation formats. By selecting such standards, agencies will be able to reduce the frequency of data reformatting required to migrate data through different standards and technology and thus to minimize the cost of digital image data preservation.

3 Methodology

The Office of the Assistant Secretary of Defense Command, Control, Communications and Intelligence (OASD/C3I) awarded a Task Order (Imaging Standard Policy Support, GS-35F-4863G/GA22) to Lockheed Martin in June 1998. The Task Order was for support in DoD's continuing study of digital image standards and the identification of the most appropriate digital imaging standard for long-term preservation of Federal documents.

With a focus on three specific categories of documents that have traditionally been transferred to NARA for long-term preservation: personnel records, manuals, standards, directive type material, and documents scheduled for declassification , the study included:

- research, gathering information from technical literature, and interviews on the status of digital imaging standards
- conduction of a survey to determine the volume, quantity, and format of electronic images and standards that each DOD activity will store and retrieve from its own libraries or transfer to the National Archives for long term preservation.
- facilitation of DOD/NARA sponsored consensus-gathering meetings and workshops; and
- publication of recommendations and findings on imagery standards for electronic records.

In October 1998, following initial research, literature review, and interviews, a survey focusing on the current usage of electronic images and archives was sent to thirty-five DOD and other Federal agencies. The survey was disseminated and returned via Email. The results were tabulated using an Access database. A copy of the survey and results are provided at the end of this report in Appendix B.

The DoD-NARA Scanned Images Standards Conference, was held March 31- April 1, 1999, at the National Archives in College Park, Md. The conference objective was to facilitate a joint Government, academic, and industry environment which would incorporate survey findings with technical knowledge and experience to determine optimum recommendations for DOD and NARA. It was attended by over 90 individuals eager for an opportunity to learn and exchange information on the current status of imaging in DoD and other Federal agencies. The Program included an overview of imaging standards including the types and extent of their use, and the status of selected imaging projects and standards associated with imaging. A summary of the conference can be found in Appendix E of this report.

In conducting the survey and facilitating the workshops, Lockheed Martin updated and expanded on the information determined in the previous DOD studies on imaging standards. Contacts with government, selected industry, and academia representatives were initiated for the survey as well as workshop attendance. OSD/C3I and NARA personnel were keep informed of study progress and findings on a regular basis via e-mail and monthly status meetings, Deliverables included both a preliminary report, which was distributed to conference attendees, this final report; monthly status reports, a task order management plan; and an action plan.

4 Results and Discussion

4.1 Current Status of the Three Focus Areas

4.1.1 Personnel Records

The Official Military Personnel Files (OMPF) include active duty health records, clinical records and medical treatment records. There are four major categories of personnel material:

- Service computation (enlistment, extensions, discharge)
- Professional History (education, training, promotions, security)
- Performance (performance evaluations, photographs)
- Administrative (dependent data, medical, loans, tuition assistance)

All of the Services have converted or are converting their personnel records to digital images in the TIFF 4 format, but have not utilized common indexing and system architecture. Therefore, while all these records are in TIFF format the header information has been entered differently. The Defense Personnel Records Imaging System (DPRIS) is an OSD initiative to move toward a common operating environment for electronically querying Official Military Personnel File (OMPF) records systems. DPRIS employs Web technologies to support electronic queries of the disparate OMPF systems and speed up search response times. Since the OMPF plan ultimately is to go entirely to database records, the number of TIFF records will eventually become stable.

Military Personnel Records (MPR) for discharged and deceased veterans are maintained at the National Personnel Records Center (NPRC) in St. Lousi, MO. Records are usually transferred to NPRC within six months after discharge or death.

4.1.2 Manuals, Standards, Directive Type Material

This category of documents have traditionally been published and distributed in paper form. DoD recently decided to stop paper publication and to make all dissemination electronically via the web, allowing ease of access to the most recent version and the ability to print on demand when paper copy is required. These records are in Microsoft Word, Hypertext Mark-up Language (HTML), Standard Generalized Mark-up Language (SGML), and Portable Document Format (PDF) formats. They will have text and embedded pictures and graphics.

The Army Logistics Support Activity (LOGSA) a reports on their web page that “Paper Technical Manuals, going...going...gone..!” LOGSA is in the process of converting paper technical manuals to CD-ROM digital media. These Electronic Technical Manuals (ETMs) are, according to LOGSA, more efficient to use and will substantially reduce deployment loads. The CD-ROMs are configured by weapon system and commodity groups with the information "tagged" to link the user with corresponding information and drawings within the document. The ETMs will be distributed using the U.S. Army Publishing Agency (USAPA) system. All conversion will be completed by the end of FY 1998 with sustainment beginning in FY 1999.

The Defense Automated Printing Service encourages the use of digital files. They identify the following reasons for using digital files:

- Reduces “hidden” cost of printing
- Reduces obsolescence
- Reduces storage costs
- Reduces transportation costs
- Allows documents to be “where you want them, when you want them”
- Every printed copy is an original – produced at maximum resolution of the print device
- Allows the captured data to be used for more than one purpose

- Compresses document cycle time from “author” to “user”
- Streamlines business process
- Increases customer satisfaction

4.1.3 Declassified Documents

Executive Order 12958, signed by President William J. Clinton on April 17, 1995, mandates that all Executive Branch records of historical interest that are older than 1976 and that are classified as National Security Information be reviewed and, with the exception of nine basic exemption categories, be declassified and made available to the public by April 2000. The number of pages in this category government-wide is uncertain but is estimated to be about 2 billion.

Since many of the records must be reviewed by several agencies it was decided to use a digitized image of the record for redaction. The document declassified project has chosen the TIFF 6 format. Committing the classified documents to digital form has allowed for greater ease in exchange of the documents for review and redaction when more than one agency is required to review the document for declassification.

The Electronic Document Interchange Standard (EDIS) is a voluntary standard for electronic document interchange among Executive Branch agencies, which review electronic images of documents. The standard governs both document metadata and document images that are to be exchanged for purposes of coordinating review, as well as minimum transfer metadata. This Standard is designed solely to provide specifications for the interchange of electronic documents and related information between systems. The Standard was developed by the Declassification program Managers Council (DPMC) Automation Working Group (AWG) and The George Washington University Declassification Productivity Research Center (DPRC) for the declassification community.

Once these 2 billion records have been declassified, they will be destined for the National Archives. This process to be carried out annually from now on, as documents reach the their expected time for declassification.

4.2 Electronic Record Imaging for Preservation

Agency records are being produced as electronic documents at an ever-increasing rate. As noted earlier in section 2.2, many government organizations are moving away from the traditional process of producing and storing paper hard copies of documents, and are moving into the realm of maintaining documents electronically.

In order to avoid the prospect of these records becoming obsolete and unreadable, there are a number of issues associated with electronic record processing that must be addressed and dealt with. The primary purpose is to ensure these electronic records can be used at any time in the future.

The file format that the electronic records are stored in has to remain readable in the future, in order to ensure these electronic records remain useable. Currently, electronic records of documents are produced in several different ways, and each method has varying levels of risk associated with it.

The most common method of archiving electronic records is to simply store the electronic file that was generated when the document was created. This application specific format could be in ASCII text format, Microsoft Word format(s), Word Perfect format(s), Microsoft PowerPoint format, or any one of hundreds of different applications currently in use by the Federal Government today. This is the most cost-effective method of storing electronic records of documents, but possesses the highest risk of the document being lost or becoming unusable.

In today’s rapidly changing technology world, the application vendors need to be constantly changing their products to keep up with their competitors. The product life cycle, from the introduction of a new version of the application to the release of the next version, is typically between 1 year and 18 months. For the most part, the newer application will maintain backward compatibility with the older version, meaning that files generated with the older version of the product can be viewed, edited and printed with the newer version. However, a 100% compatibility between two subsequent versions of the same product can never be guaranteed. In order to ensure that the documents can be reproduced in the newest version of the application, the document needs to be opened in the newer version, and a quality assurance check needs to be performed. Typical errors that are discovered when a new release of the product becomes available is that embedded information in the document, such as page numbering,

placement of graphics, table formatting, heading numbering etc. are lost or changed. These items have to be corrected, and the document saved in the new format in order to ensure the electronic record of the document remains useable in the new format.

A preservation option for storing electronic records is to create either a vector or raster electronic image of the document. This can be accomplished by scanning the original document and saving the resultant file, or by converting the application file directly into a graphic image file.

Both of these options produce an electronic image of the original document. The method chosen will depend on the organization's preferences and budget. The primary problem associated with these methods of creating a electronic file suitable for archival and long term storage of the resultant electronic file is that a fairly large expenditure of time and resources is typically required in comparison with just storing the application file. This method of producing an electronic record is typically not accomplished for every record produced within a government organization, and due to the cost involved should not be accomplished. The government organization has to make a decision on which documents need to be preserved, how long they need to be preserved for, and how many resources need to be expended to preserve the documents. This process in itself adds to the cost of preserving an electronic record of the document.

The file format that is used store the image of the document is, or should be, a primary consideration when deciding to create an electronic image of the document. There are a number of different image formats currently available for this purpose, and each one has its advantages and disadvantages. The following section describes some of the most widely used imaging formats, and the advantages and disadvantages of each.

4.3 Data Format Standards

Data format standards that have received the approval or endorsement of a standards body such as the American National Standards Institute (ANSI) or the International Organization for Standardization (ISO) are referred to as de jure standards. On the other hand, data format standards that become a standard by sheer volume of usage and acceptance by users are called de facto standards. In the digital imaging arena both types of standards have advantages and disadvantages, and all carry a certain level of risk.

De jure standards take a long time to develop and must be approved by every organization that is a member of the standards organization with interests in the area covered. They are developed and maintained by a group or board of professionals. Suggested changes and updates to the standard are carefully reviewed according to a controlled process. These standards tend to be broad in scope. Frequently de jure standards are considered cumbersome and restrictive. The biggest risk associated with de jure standards is that they will never achieve user acceptance and industry penetration. Examples of de jure standards include: US MARC, JPEG, Z39.50, BIIF (Basic Image Interchange Format), and SGML.

De facto standards spring up in response to an immediate industry need. They gain in use and popularity at the dictates of the market. They are usually maintained by the group or business that originated the standard, with no community review. These standards tend to be narrower in scope and designed for one specific purpose. They penetrate the market and become a standard by virtue of the fact that that is what is available. De facto standards are a high risk choice for those looking for long term programs. De facto standards are not rigorously enforced. Several different, incompatible versions of one standard may exist at any given time. De facto standards are generally the proprietary property of one company or organization. De facto standards migrate and change very rapidly based on the needs of the information technology (IT) user community, which can result in a de facto standard becoming obsolete in a very short period of time. Since the de facto standards are proprietary property, the availability of the standard cannot be guaranteed for any great length of time. The company that holds the rights to the de facto standard may collapse or be taken over by another organization that does not support the same de facto standard. The economy and the financial stability of the company controlling the de facto standard play a large role in how long the de facto standard will be available for use within the IT community. Examples of de facto standards include the Tagged Image File Format (TIFF) and the Portable Document Format (PDF).

A survey was sent to various Government agencies in October 1998 to determine which electronic formats were currently being used to generate digital images of documents (see Appendix B). The top six responses to this

inquiry were either de facto standards, proprietary file formats, or 'unofficial' forms of approved standards (HTML is a form of SGML).

The tables in Appendix A identify and consolidate information about the most common formats for imaging. The first table contains those in the raster format with vector orientated formats identified in the second table.

4.3.1 Digital Imaging File Formats

The choice of the file format used to create a digital image is critical for supporting the main function of an electronic archive, that is allowing a document that is created today to be retrieved and used at any time in the future. If the file format that is used to generate the electronic image is not supported at the time the person wishes to access the document, then chances are very high that the document file will be unreadable and unusable. For this reason, the organization maintaining the electronic archive needs to make sure that the imaging file formats for the documents contained in the electronic archive are current and up to date.

Each imaging file format has a life cycle of its own, from the development and release of the format from the developer to the stage where the format is no longer supported by the commercial industry. This life cycle for imaging file formats is illustrated in the following figure. The life cycle runs from technology innovation to obsolescence. The most common imaging file formats are shown on the graph, representing the appropriate life cycle phase for each of the most commonly used file formats.

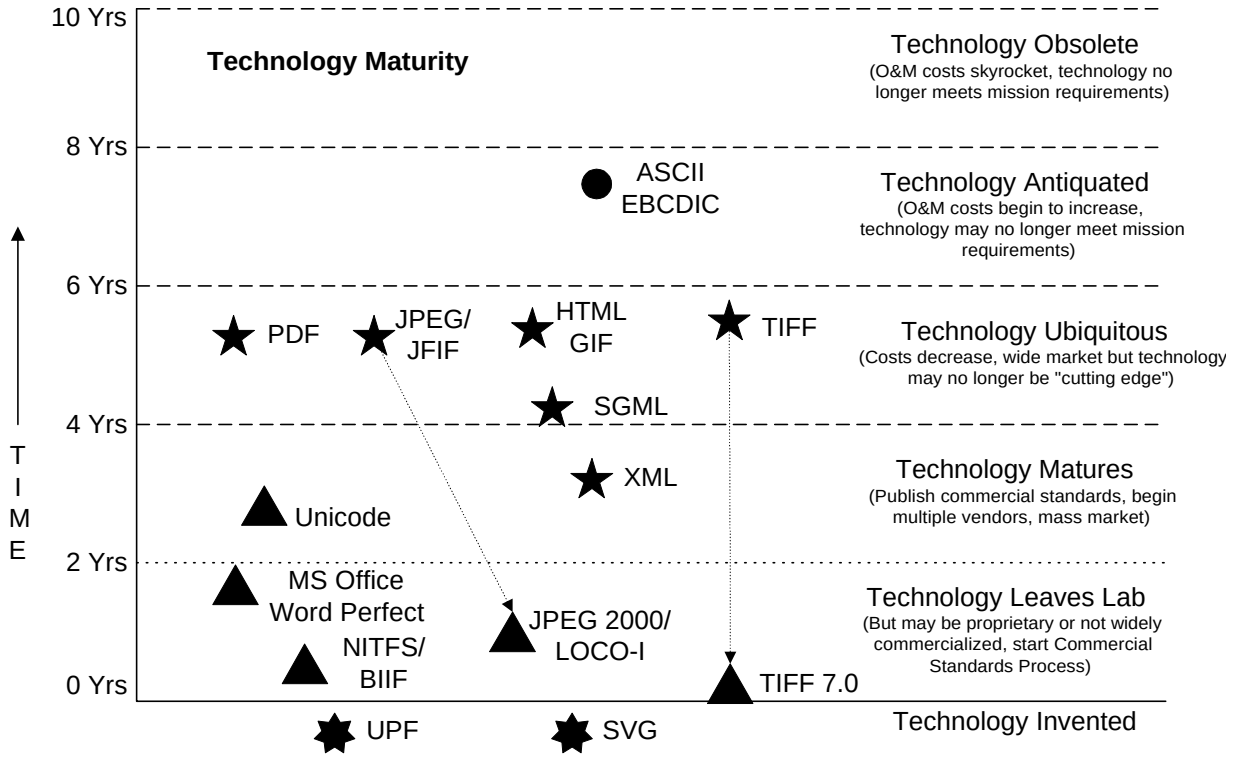

FIGURE 1. TECHNOLOGY MATURITY

This diagram also represents the maintenance costs associated with each imaging file format depending on the life cycle phase and the current migration paths for some of the formats. There are a several new file formats that are currently just over the horizon with the promise to replace the existing file formats, if there is a wide enough acceptance in the commercial marketplace. Some of the file formats listed in the figure are described in more detail in Section 4.2, based on the current popularity, acceptance and usage of the file format.

One of the most significant costs associated with the life-cycle maintenance of an electronic archive can be the migration cost of moving a document from one file format to another. The effort associated with changing the file

format can be as simple as opening the document and saving it as the new format. However, experience has shown that migration is usually not this simple. When a new file format becomes available, the manufacture usually attempts to make sure that previous versions of the format are fully supported by the new format. However, there always seems to be something that doesn't work correctly, and someone has to spend a significant effort in reformatting the document file to make the new version look like the original.

An example is opening a document that was created in a previous version of Microsoft Word for Windows. Problems are usually encountered with the page layout, heading numbering, graphics, etc. even for documents that were created using the previous revision. Opening documents that were created in earlier revisions creates even more problems. This means that someone has to spend the time to reformat the document in the new format to make sure that it looks identical to the original document. If there is not a hard copy of the document available and the user does not have a working copy of the previous version of the software, then it can be extremely difficult to reformat the document to look exactly like the original.

With documents that contain only text, this much less of a problem. However, with the technology available today, word processors that make it easy to add all sorts of graphics to a document. This makes the presentation and layout of the document as important as the text itself. More and more people are becoming reliant on the old adage 'that a picture is worth a thousand words'. If the picture that represents the thousand words cannot be viewed or located when the document is opened in the future, the document itself looses a significant part of its meaning.

This is not to say that the documents created today should not contain any graphics or pictures. What this means is that the organizations that wish to create and maintain an electronic file of the document as an archived record, they need to take this into consideration when selecting the format for archiving and preserving the document.

4.3.2 File Format Usage in Government Organizations

In the survey of the Government organizations previously mentioned, of the organizations that currently maintain electronic document archives provided the following results when asked which file formats were currently being archived:

- 100% of the organizations archive TIFF formatted documents.
- 73% of the organizations archive PDF formatted documents.
- 55% of the organizations archive Joint Photographic Experts Group (JPEG) formatted documents.
- 45% of the organizations archive Graphics Interchange Format (GIF) formatted documents.
- 45% of the organizations archive HTLM formatted documents.
- 27% of the organizations archive SGML formatted documents.
- 27% of the organizations archive Microsoft Word (.doc) formatted documents.
- 18% of the organizations archive text (.txt) formatted documents.
- 18% of the organizations archive ASCII formatted documents.
- 18% of the organizations archive Excel spreadsheet (.xls) formatted documents.
- 9% of the organizations archive PostScript formatted documents.
- 9% of the organizations archive Continuous Acquisition and Life-Cycle Support (CALS) formatted documents.
- 9% of the organizations archive Microsoft PowerPoint (.ptp) formatted documents.
- 9% of the organizations archive Word Perfect (.wpd) formatted documents.

The following sections describe the benefits and disadvantages of using the most popular file format types listed above.

4.3.2.1 Tag Image File Format (TIFF)

TIFF is a format standard that was developed by Aldus Corporation and Microsoft in the late 1980's as a file format designed to promote the interchange of digital image data. Since that time, there have been two major revisions to the original specification, TIFF 5.0 and TIFF 6.0. Recently, Adobe Corporation, which merged with Aldus

Corporation and assumed the rights to TIFF, has announced that the TIFF 7.0 specification will be released in the near future.

As shown by the survey results, TIFF is the most common format for storing digital images of documents. The reason for this is that almost every scanner on the market today is capable of creating a TIFF file that is an exact replica of the scanned document. As shown in the survey, 91% of the organizations that maintain an electronic archive use scanning technology to create electronic images of paper documents.

One of the major advantages to TIFF is also one of it's greatest liabilities when it comes to electronic archives. The TIFF file format is very flexible and loosely defined, and can be customized by the image's creator to support any number of functions such as compression, pallet colors, etc. What this really means is that not all TIFF viewers are capable of viewing all TIFF images. An example of this is that WordPerfect V5.x and V6 for IBM PC will read all base formats for TIFF, but will not read compressed TIFF files. Another indication of the problems encountered with TIFF files are the many entries in the Adobe (the maintainer of the TIFF specification) Technical Solutions Database describing problems people have encountered using the TIFF file format (http://www.adobe.com/supportservice/custsupport/database.html, search on TIFF).

Adobe Corporation has not published the TIFF 7.0 specification as of the date of this document.

The costs for developing a TIFF image are minimal, and coming down steadily with the price for scanners. In the last year, the cost for a high quality scanner has dropped significantly, allowing almost every organization the luxury of owning a scanner. The scanners typically come bundled with imaging software, allowing for even more options in creating a digital image.

4.3.2.2 Portable Document Format (PDF)

Adobe Corporation's PDF format has become a de facto standard for publishing documents on the World Wide Web (WWW). The PDF format has several key benefits, the most significant of which is that it is a completely device independent page description language. This open systems approach has made for a wide acceptance of PDF as the standard for publishing documents either on the web or for printed documents.

The PDF file is typically much smaller than the original document format, therefore enabling more documents to be stored on the same media. Depending on the fonts used in the original document, the PDF format may produce an exact replica of the original document, including graphics, pictures, and tables.

One of the biggest disadvantages to PDF is that it is a proprietary format that is owned by the Adobe Corporation. However, Adobe has made the PDF standards available to other vendors, and other software manufactures have created products that produce and read PDF files. The following list identifies several vendors that market PDF creation products:

- ZEON Corporation's DocuMaker program will convert any document that is saved in a postscript format to a PDF document.
- FastIO Systems provides the ClibPDF program, an ANSI C Source Library for direct PDF generation without relying on any Adobe Acrobat tools and related products.
- 5D is a company that provides the NIKNAK software tool that converts postscript files to PDF files.
- Adobe Corporation also provides a freeware program, PDFMaker for Microsoft Word 97 that works with Adobe Acrobat 3.0 for Windows to convert Microsoft Word documents into PDF files.

While there are a number of companies that provide product support for PDF, the format is still proprietary. This makes users dependent on Adobe if they want the latest product line that supports the generation of PDF file.

4.3.2.3 Joint Photographic Experts Group (JPEG)

A number of organizations reported on the survey that they were archiving documents in the JPEG format. While there is a .jpg file name extension for files using the JPEG compression method, there is not an actual file format called JPEG. There are actually at least three different file formats that use the .jpg file name extension:

- Still Picture Image File Format (SPIFF) is the official ISO standard JPEG file format.

- JPEG File Interchange Format (JFIF) is the de facto standard for JPEG images developed by C-Cube Microsystems, because it took the ISO took over five years to develop the SPIFF standard.
- Image JPEG (IMJ) was created by Pegasus Image Corporation as a variation of the JFIF file format. IMJ is essentially a JFIF file with a Microsoft Windows Bitmap (BMP) header and enhanced palette optimization. The IMJ format is used in several screensaver applications and by organizations such as Delrina and the National Center for Missing Children.

These three file formats are for the most part compatible and most JPEG readers will read all three formats. However, this is not always the case. Some JPEG readers will only open JFIF images, while still others generate an error message when attempting to open a JPEG image other than SPIFF. This problem relates to the JPEG Standard itself, which has 44 different modes for compressing images. Most of these modes are application specific.

The ISO is in the process of developing a new image compression standard called JPEG 2000. This new standard is being developed to compliment, not to replace the current JPEG standard (ISO 10918-1, ISO 10918-2, ISO 10918-3). One of the goals of the new standard is to develop a single decompression architecture that will encompass all of the different compression modes.

The baseline JPEG is classified as a lossy compression algorithm because the decompressed output is not bit-for-bit identical to the original input. The baseline JPEG compression ratio can be set to provide an output image that is visually indistinguishable from the original, but there will always be some loss of image quality. The JPEG Standard ISO 10918-3 currently contains a lossless compression algorithm, and another lossless algorithm, JPEG-LS, is in the final draft international standard FDIS14495-1.

The compressed JPEG images, since they are considered to be lossy, should not be used as the archived version of a document or image. The document that is maintained in the archive should be either the original document, or the document in a file format that is identical to the original document.

4.3.2.4 Standard Generalized Mark-up Language (SGML), Hypertext Mark-up Language (HTML), and eXtensible Mark-up Language (XML)

SGML, HTML and XML are all markup languages that were designed for the transmission of information from one computer to another. The differences between the three are quite distinct, but the basic format for the format files themselves remains the same.

All markup language files can be viewed using a standard text editor. The codes that are placed in the SGML, HTML or XML files that describe the formatting characteristics for the document are simple text codes placed in brackets. The actual information that is contained in the format file is stored as text data. Images are inserted into the file as hyperlinks: the actual files for the pictures, images and graphs displayed with the text data are not actually stored inside the markup language file. The hyperlinks provide the data path to the image or picture that needs to be inserted on the page.

On February 11, 1999 the World Wide Web Consortium (W3C) released the first working draft of the Scalable Vector Graphics (SVG) specification. This specification will allow vector graphics to be inserted as text information directly into the mark-up language file. There are several significant advantages to this new specification, the most critical is that this specification will eliminate the necessity for having more than one file. Another critical advantage to the SVG specification is that text searches can be performed on the text information contained in the vector graphic. Currently, separate metadata information about the contents of the vector graphics file must be provided if the user needs to perform a search on the image.

SGML is defined in ISO Standard 8879:1986, and is a formal language used to pass information about the component parts of a document from one computer system to another. The markups provided by SGML tell the computer that is displaying the document more than just how the information is to be displayed on the monitor, such as where it is displayed on the screen, which fonts to use, and where graphics should be inserted in the text. SGML provides a method for describing the relationships between different parts of a document, such as paragraph numbering, table of contents, indexes, etc. SGML also allows users to include metadata about the document such as the author's name, date published, etc. within the SGML file.

SGML is currently the archival imaging format of choice for many libraries because it allows users to perform a search on the text contained in the SGML file. The SGML file itself can be read and searched using a standard text editor, unlike the other imaging formats which require special software to be used such as optical character recognition (OCR) software.

HTML is an application of the SGML that uses a predefined set of document type definitions (DTDs) that are used to markup documents, describing how the document should be formatted for the user's screen. The difference between HTML and SGML is that SGML does not provide a standard set of DTDs. The document's creator can define the DTDs, and passed along with the SGML file to the computer that requests the file over the Internet.

XML is a subset of the SGML standard that is becoming more and more popular, and may one-day replace both HTML and PDF as the most prevalent web publishing formats. The difference between HTML and XML is that XML allows users to specify their own customized tags the same as with SGML, but is not possible with HTML. This capability, of letting the document writer prepare and provide their own DTD, creates an extra file that the web browser has to download from the source site to determine the meaning of the customized tags in the document.

The advantages that are available with SGML and XML that are not available with HTML are in the area of metadata. With SGML and XML, author of the documents is able to insert metadata such as the author's name, the date published, and the topic or subject of the document. The metadata can be marked with custom tags, such as <author/> or <subject/> that allows users to search for the document using this criteria. With HTML, this type of information needs to be included in with the text itself, rather than as metadata or 'data about the data.'

4.3.2.5 Scalable Vector Graphics (SVG)

The first working draft of the SVG specification was released by the World Wide Web Consortium (W3C) on 11 February 1999. This file format specification promises to change the way that vector graphics are inserted into the Markup Language file formats.

With the current Markup Language file formats, each graphic is contained in its own separate file. The HTLM, SGML or XML document must contain a hyperlink to the graphics file. This results in the document creator having to maintain, update, and edit several different files for each document. This also means that when an electronic document is placed in the electronic archive, all graphics and text files must be present. The hyperlinks contained in the markup language document must be updated to point to the correct location, which might change every time the document is moved from one physical location or media type to another. This also prevents the document user from being able to perform searches on the text contained in the graphic file.

The SVG specification changes all of this. With SVG, the vector graphic is inserted directly into the Markup Language document eliminating the need for the hyperlink to a separate file. A second major advantage to the SVG specification is that the text contained in the vector graphic now becomes a part of the main document, allowing user's to perform searches on the text.

Most of the major graphics software vendors, including Adobe, Apple, Autodesk, Corel, HP, IBM, Inso, Macromedia, Microsoft, Netscape, Quark, RAL, Sun, and Visio have been supporting the development of the SVG specification. This indicates that there will be wide industry acceptance for this new graphics format, and promises to change the way HTML, SGML and XML documents are generated and archived.

4.3.2.6 Universal Preservation Format (UPF)

In 1996, the National Historical Publications and Records Commission of the National Archives awarded a grant to WGBH, a public broadcasting station in Boston Massachusetts, to research and produce a prototype of a platform-independent Universal Preservation Format (UPF). This file format would be designed specifically for digital technologies that will ensure the accessibility of a wide array of data types into the indefinite future. A draft document describing this initiative can be found on the WWW at http://info.wgbh.org/upf/.

4.3.3 Establishing and Maintaining an Electronic Archive

There are a number of methods currently in use today for the generation of digital images of documents to be stored in an electronic archive. The choices taken by the organization that is responsible for the preservation of the document are dependent on a number of factors, cost usually being one of the most critical criteria.

The cost of creating and maintaining an electronic archive is much greater than just the cost of creating the digital image, storing the resulting electronic file on some type of media, and placing the media in a safe location. The long-term costs such as migration of records from one format or media to another must be taken into account or the organization runs the risk of not being able to retrieve documents from the archive at a later date.

The consideration in creating an electronic archive is determining which documents generated within the organization need to be preserved, and for how long these documents need to be preserved. As shown in the survey results, this will vary from organization to organization. While a majority of the organizations reported that less than 10% of the documents stored in an electronic archive need to be preserved for a long period of time, several organizations reported that up to 100% of the records stored in their electronic archive need to be permanently preserved at the National Archives. (see Appendix B)

Being able to access records stored in an electronic archive entails much more than just storing the electronic files on a network drive, CD-ROM, or optical drive. The records themselves need to be cataloged, indexed, and linked to a text file that provides an explanation of the contents of the image. If this is not accomplished, then while the records themselves may be preserved, the information contained within will not be very useful to others trying to access the records in the future.

This information about the image or electronic record is referred to as metadata.

To be truly efficient, an electronic archive should be built on a database concept, where the image can be linked to the metadata text file and other information supporting the image.

There are three factors which are critical for ensuring an electronic archive is established that minimizes the life cycle costs and ensures that the information contained in the archive can be retrieved at any time it is required:

- The format that is used to create the digital image
- The media that is used to store the digital images, and
- The use of a Records Management System (RMS) or a Document Management System (DMS),which is DoD 5015.2 compliant, to manage the electronic archive.

4.4 Metadata

Metadata is data about data. In this case data or information about the image or electronic record. Metadata typically support a specific function: discovery or access; administrative; or structural. Access metadata include location, subject, authors, creator, etc. Administrative metadata include type of item, file format, compression format, dimensions, bit-depth, color lookup table, etc. Structural takes administrative data one step further and identifies file size relationship to other file records.

The standard for bibliographic data is US MARC. In its complete format it is designed to be a transfer format of bibliographic data from one system to another. Many feel that MARC records are too expensive and time consuming. However, it is not necessary to do a complete AACR2 (Anglo-American Cataloguing Rules, Second Edition) catalog record to have a MARC record. You merely need to identify your selected fields or tags with the associated MARC field identifier and your record will be accessible on thousands of Commercial-Off-The-Shelf (COTS) products designed to search or access data.

The desire for increased access to electronic records and to the Web has driven initiatives such as the Encoded Archival Descripting (EAD), Dublin Core, the Text Encoding Initiative (TEI) and the Resource Definition Framework (RDF) Extensible Markup Language (XML). Every electronic file format must contain some form of metadata to tell the computer how to display the record. This is usually called the header of the record and thus there is a TIFF metadata standard and a BIIF metadata standard, etc. Specialized collections of data have also created metadata standards. For example the Federal Geospacial Data Committee (FGDC) has established a metadata set for geo-spatial data (digital maps and related items). There is also the Warwick Framework, an

architecture that allows for the interchange of distinct metadata packages, Z39.50, and the set of required metadata found in DoD 5015.2 Std.

Each format has structure and administrative data in its header information. Ideally this data should be standardized. However, front-end search can work through a defined set of differences. Example of this is the front-end work to provide access to Official Military Personnel Files (OMPF) records.

The table that follows is a sampling of fields from some of the most common metadata standards. It illustrates that both a core of data can be found and that specialized fields are required for different categories of images.

TABLE 1. COMMON METADATA STANDARDS

DoD 5015.2 - STD	Dublin Core	MARC	GILS	EAD
Subject	Subject/keywords	650, 653$a	Uncontrolled term	
	Title	245$a	Title	<titleproper>, <unittitle>
Author or originator	Author/Creator	700, 710, 720$a	originator	<origination>
Originating organization	Publisher	260$b	Distributor	<publisher>
	Other Contributor	720$a, 700, 710	Contributor	
Document creation date	Date	260$c	Date of publication	<date>, <unitdate>
Media type	Resource Type	655$a	medium	
format	Format	856$q	Available linkage type	<physdesc>
	Resource Identifier	856$u	Available linkage	
	Relation	787$n	Cross reference relationship	
	Source	786$n	Sources of Data	
	Language	546$a , 041$a	Language of Resource	
	Coverage	500$a, 255$c, 513$b	Supplement information Bounding coordinates Time period textual	
	Rights	540$a	Use constraints	
	description	520$a	abstract	<abstract>
Date filed		X		
Addressee				
Location of record		X		

DoD 5015.2 - STD	Dublin Core	MARC	GILS	EAD
Vital record indicator				

4.5 Costs

The cost of preparing electronic files or records for archiving documents is a critical factor in the decision making process. However, due to a number of factors, the costs associated with the long-term preservation of electronic documents are extremely complex and very fluid. The difficulty associated with creating a cost model for archiving documents has been addressed by a number of researchers and professionals working in the electronic archiving field, but a complete cost analysis has never, to the authors knowledge, been published. At best, any cost model can only be a snapshot of the archiving costs for any given time.

The following quote illustrating this point is taken from the Cornell University Digital to Microfilm Conversion: A Demonstration Project 1994-1996, Final Report to the National Endowment for the Humanities by Anne R. Kenney:

> "Numerous conferences and reports have been dedicated to issues associated with digital archiving-ensuring continuing access to digital materials across hardware/software configurations and subsequent generations of computer technology. The clearest articulation of these issues is provided in the Joint Task Force Report of the Research Libraries Group and the Commission on Preservation and Access, entitled *Preserving Digital Information: Final Report and Recommendations*. As the report makes clear, currently there are no agreed-upon processes or model institutional programs for preserving digital collections over time. **There is even less consensus on the costs of such efforts.**"

Some of the reasons for the complexity of the cost model are as follows:

- The information technology field is the most rapidly changing industry today. Almost as soon as a new product or standard becomes available on the market, it becomes obsolete due to new advances in technology, new products that hit the market and new requirements that are being identified due to these new technologies and products.
- The costs for information technology are constantly fluctuating, not only decreasing in some areas but increasing in others as well.

There are a number of factors that must be looked at when determining the total costs associated with archiving electronic documents. The costs for creating the electronic file or record are not the only costs that must be looked at. The person or group preparing the electronic files for archiving purpose needs to look at the entire lifecycle of the document or record. This not only includes the short term archival of documents at the preparer's facility, but also the cost associated with maintaining the documents at long-term storage facilities such as the National Archives or a data warehouse.

The factors that must be considered are as follows:

- Selecting documents for storage.
- Short-term document storage (at preparer's facility).
- Cost of transferring the document to a long-term storage facility, to include document conversion from one format to another, the media that will be used to transport the document, etc.
- Costs of maintaining the document in the long-term storage facility.

This short list assumes that the organization has already developed standards, methods and processes for creating electronic images, storing electronic files, and transferring electronic records between organizations. If these efforts have not already been accomplished by the organization responsible for the life cycle of the document, then the costs associated with these efforts must be factored in as well.

During the course of our research we collected data from organizations that market the service of digital image record creation. The following table provides cost information that was obtained from one organization that is

currently involved with a large-scale document-imaging project that is creating electronic files for a government organization:

TABLE 2. COST INFORMATION

Method of Conversion	Format	Price
Single pages to electronic image	TIFF	10-16¢ /page
Bound book pages to electronic image	TIFF	22-25¢/page
Film image to electronic image	TIFF	10¢/page
Fiche image to electronic image	TIFF	12¢/page
Single page or from TIFF image	PDF with text at 96% OCR accuracy	20¢/page - $2.10/page

Throwing out the high price (labor intensive – high percentage of text re-keyed textual information) the comparison is provided in the following graph:

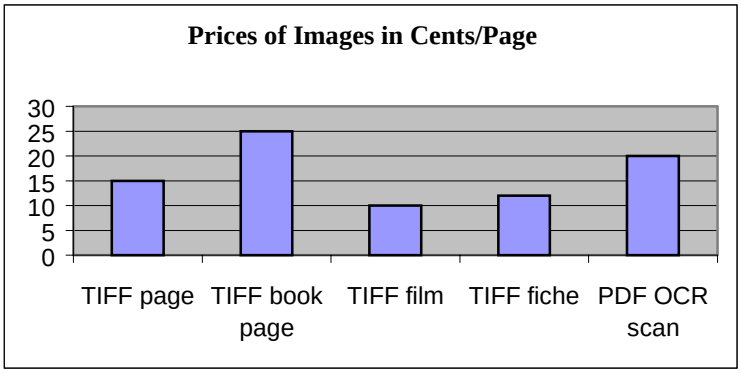

FIGURE 2. IMAGE PRICES

These low-end prices are for material that does not require a lot of preparation and indexing and material that does not require special handling and can be feed automatically into the imaging equipment. Equipment costs, acquisition and maintenance, staff training, quality control, verification, metadata creation, and software costs are spread over a large business base.

Imaging from microfilm is almost entirely automated and has lower associated costs since the preparation has been previously done when material originally microfilmed. For projects that are intended for both archive and access, the industry recommends producing not only the scanned TIFF image, but in the same process, feed the image to a computer-output microfilm (COM) machine and produce micofilm for archival purposes. The rule of thumb is that if you want ***access,*** use the electronic image, but if you want ***preservation,*** then use microform.

A second cost sample was obtained from the final report published by Cornell University on the *Digital to Microfilm Conversion: A Demonstration Project*. Cost figures were provided in this report not only for the conversion of digital images to microfilm, but also for the conversion of microfilm to digital images. Yale University (Project Open Book) conducted the microfilm-to-digital project at the same time as the Cornell University project, and information was shared between the two institutions. The following cost information was provided in the *Cornell University Digital to Microfilm Conversion: A Demonstration Project 1994-1996, Final Report* to the National Endowment for the Humanities by Anne R. Kenney:

TABLE 3. PRODUCING DIGITAL IMAGES FROM PAPER VS. MICROFILM

	Cornell: Time & Costs			Yale: Time & Costs		
Process	Mean Time	$/Bk	$/Image	Mean Time	$/Bk	$/Image
Preparation	78.8 min	$20.20	$0.094	5.3 min	$1.36	$0.006
Scanning						
Auto	56.1 min	$14.38	$0.067	38.1 min	$9.77	$0.045
Manual	73.2 min	$18.76	$0.087			
Indexing	8.6 min	$2.20	$0.010	29.9 min	$7.66	$0.035
Other	5.2 min	$1.33	$0.006	19.2 min	$4.92	$0.023
Sub Total: Process						
Auto	148.7 min	$38.11	$0.18	92.5 min	$23.71	$0.110
Manual	165.8 min	$42.49	$.020			
Equipment	Mode			Capacity		
	Auto	$14.30	$0.066	High	$24.51	$0.113
	Manual	$17.40	$0.080	Low	$31.32	$0.145
Total: Process/Equip	Mode			Capacity		
	Auto	$52.41	$0.24	High	$48.22	$0.22
	Manual	$59.89	$0.28	Low	$55.03	$0.26

The following table provides direct comparison of the two data samples:

TABLE 4. DIRECT COMPARISON

	Sample 1	**Sample 2**
Digital Image per Page	$0.13[1]	$0.25[2]
Digital Image per Bound Book[3]	$54.00	$53.88[2]

Notes:

1. This was the average cost reported in Sample 1, which ranged from $0.10 -$0.16/page.
2. These figures are an average of the Cornell University and Yale University costs, which ranged from $0.22 to $0.28.
3. Both figures are based on a 216-page book.

There are several factors that can explain the difference in the Digital Image per Page cost between these two sets of data. One is that the Cornell and Yale University studies were established as demonstration or proof-of-concept projects, while the Government project was a competitive bid effort. Another explanation could be that the University projects were accomplished in-house, while the Government project was contracted out. Regardless of

the differences, these examples serve the purpose of providing cost information that can be used for planning purposes. For more detailed information, the Cornell University report can be downloaded from the WWW at http://www.library.cornell.edu/preservation/com/comfin.html.

Within in the DoD the vendor or outsourcing agent is the Defense Automated Printing Service (DAPS). DAPS is responsible for document automation and printing within the Department of Defense, encompassing electronic conversion, retrieval, output and distribution of digital and hardcopy information.

DAPS sees conversion as one of the most important services to their customers over the next few years. The DoD has issued strategic goals and objectives that require the DoD to transition into paperless environments. The DoD is looking to DAPS for support in moving from a paper-based to a digital environment, including raster scanning, engineering drawings, Tag Image File Format (TIFF), Group 4 format, quality Assurance, CD-ROM and WORM, SGML to HTML, hyperlinked PDF and much more.

4.6 Migration Strategies

As the operating environments of digital archives change, it becomes necessary to migrate their contents. There are a variety of migration strategies for transferring digital information from systems as they become obsolete to current hardware and software systems so that the information remains accessible and usable. No single strategy applies to all formats of digital information and none of the current preservation methods is entirely satisfactory. Migration strategies and their associated costs vary in different application environments, for different formats of digital materials, and for preserving different degrees of computation, display, and retrieval capabilities. The general rule of thumb appears to be plan for your migration efforts to cost between 50 – 100% of the cost to create the original digital image document.

Methods for migrating digital information in relatively simple files of data are quite well established, but the preservation community is only beginning to address migration of more complex digital objects. Additional research on migration is needed to test the technical feasibility of various approaches to migration, determine the costs associated with these approaches, and establish benchmarks and best practices. Although migration should become more effective as the digital preservation community gains practical experience and learns how to select appropriate and effective methods, migration remains largely experimental and provides fertile ground for research and development efforts.

One migration strategy is to transfer digital materials from less stable to more stable media. The most prevalent version of this strategy involves printing digital information on paper or recording it on microfilm.

Retaining the information in digital form by copying it onto new digital storage media may be appropriate when the information exists in a "software-independent" format as ASCII text files or as flat files with simple, uniform structures.

Copying from one medium to another has the distinct advantage of being universally available and easy to implement. It is a cost-effective strategy for preserving digital information in those cases where retaining the content is paramount, but display, indexing, and computational characteristics are not critical. As long as the preservation community lacks more robust and cost-effective migration strategies, printing to paper or film and preserving flat files will remain the preferred method of storage for many institutions and for certain formats of digital information.

Another migration strategy for digital archives with large, complex, and diverse collections of digital materials is to migrate digital objects from the great multiplicity of formats used to create digital materials to a smaller, more manageable number of standard formats that can still encode the complexity of structure and form of the original. A digital archive might accept textual documents in several commonly available commercial word processing formats or require that documents conform to standards like SGML (ISO 8879). Databases might be stored in one of several common relational database management systems, while images would conform to a tagged image file format (TIFF) and standard compression algorithms (e.g., JPEG).

Changing format as a migration strategy has the advantage of preserving more of the display, dissemination, and computational characteristics of the original object, while reducing the large variety of customized transformations that would otherwise be necessary to migrate material to future generations of technology. This strategy rests on the

assumption that software products, which are either compliant with widely adopted standards or are widely dispersed in the marketplace, are less volatile than the software market as a whole. Also, most common commercial products provide utilities for upward migration and for swapping documents, databases, and more complex objects between software systems. Nevertheless, software and standards continue to evolve so this strategy simplifies but does not eliminate the need for periodic migration or the need for analysis of the potential effects of such migration on the integrity of the digital object.

Use of one of the evolving interchange standards, such as the Basic Image Interchange Format (BIIF) or Electronic Document Interchange Standard (EDIS) allows for the receipt of images in many different formats which are converted into one robust format. Having only one format that will handle all types of images simplifies the migration issue to handling of only one format.

BIIF is based on the National Imagery Transmission Format Standard (NITFS) developed by the DoD and adopted by North Atlantic Treaty Organization (NATO). The BIIF is the basis for a new standards activity within ISO/IEC JTC1/SC24 to add a new part 5 to the International Standard for Image Processing and Interchange (IPI) (ISO 12087-5, 1998)

BIIF specification provides such a common basis for storage and interchange of images and associated data among existing and future applications. BIIF supports interoperability by providing a data format for shared imagery and an interchange format for images and associated imagery data. The documentation provides a detailed description of the overall structure of the format, as well as specification of the valid data content and format for all fields defined within a BIIF file. BIIF provides a data format container for raster, symbol, and text data, along with a mechanism for including image-related support data.

BIIF satisfies the following requirements:

- Allow diverse applications to share imagery and associated data.
- Allows an application to exchange comprehensive information to users with diverse needs or capabilities, allowing each user to select only those data items that correspond to their needs and capabilities.
- Minimizes preprocessing and post processing of data.
- Minimizes formatting overhead, particularly for those applications exchanging only a small amount of data and for bandwidth-limited systems.
- Provides a mechanism to interchange Programmer's Imaging Kernel System (PIKS) (Part 2 of ISO 12087) image and image-related objects
- Provides extensibility to accommodate future data, including objects. As BIIF becomes more capable through extension and the addition of new data, objects and data relationships, concepts and features of 12087-3 (Image Interchange Format [IIF]) may be considered as a more appropriate method of growth. This is to facilitate a growth path from BIIF to IIF.

In BIIF, data interchange between disparate systems is potentially enabled by a translation process. Using BIIF, each system must be compliant with only one external format that will be used for communication with all other participating systems. When BIIF is not used as a system's native internal format, each system will translate between the system's internal representation for imagery and the BIIF format. A system from which data is to be transferred has a translation module that accepts information structured according to the system's internal representation for images and related imagery data, and assembles this information into the BIIF format. The receiving system will reformat the BIIF data, converting it into one or more files structured as required by the internal representation of the receiving system. Each receiving system can translate selectively and permanently store only those portions of data in the received BIIF that are of interest. A system may transmit all of its data, even though some of the receiving systems may be unable to process certain elements of the data.

Profiles of BIIF will be established as International Standardized Profiles (ISP) through the ISO process (ISO/IEC TR 10000).

EDIS is a voluntary standard for electronic document interchange among Executive Branch agencies, which review electronic images of documents. The standard governs both document metadata and document images that are to be

exchanged for purposes of coordinating review, as well as minimum transfer metadata. This Standard is designed solely to provide specifications for the interchange of electronic documents and related information between systems. The Standard was developed by the Declassification Program Managers Council (DPMC) Automation Working Group (AWG) and The George Washington University Declassification Productivity Research Center (DPRC) for the declassification community.

5 Conclusions

1. Access and response to Freedom of Information Act (FOIA) requests are facilitated through electronic digitalization of records.
2. No de jure standard for digital images has reached the desired maturity level for archival purposes.
3. The hardware and software technology required for the use of digital images changes rapidly.
4. Migration costs associated with archiving of digital images of textual material are unknown.
5. The anticipated high cost associated with long-term maintenance of digital image records mandates careful screening and selection of only the most valuable digital imaged records to be accessioned into the National Archives.
6. Metadata standards have been developed, but no one standard has emerged as the most universally accepted standard for electronic image records.
7. Tag Image File Format (TIFF) and Portable Document Format (PDF), both de facto standards, are the most widely used formats for text records.
8. Joint Photographic Experts Group (JPEG), a de jure standard, is the most widely used compression standard.
9. The use of proprietary standards for producing and storing images is much more common than the use of official standards.
10. Organizations will continue to use the proprietary imaging formats due to the costs involved.
11. The key roadblock to a successful digital imaging program is the high costs associated with the program and the lack of management understanding to the need for appropriate funding in the area.
12. The lack of a format standard is no longer seen as a major issue
13. A united government voice was needed, with strong NARA leadership and a means of sharing data.

The following phased implementation approach received general acceptance at the DoD-NARA Scanned Images Standards Conference:

1. Manage the process (records management, management and policy)
2. Study, plan, gather information through cost/benefit analysis of entire life-cycle (especially document preparation, searching, and migration).
3. Pick an interim standard during step 2, which will be accepted and supported by DoD and NARA – this will enable the cost-benefit analysis to be conducted.
4. Practice migration and preservation while documents are in active use

6 Recommendations

1. Image electronic digital material in the most stable formats available preferably using the latest version, but no more than two generations prior to the latest. (e.g., for TIFF image produced in January 1999 that would be TIFF version 6, 5 or 4)
   a. Image personnel records using TIFF for archiving, TIFF or PDF formats for access. Convert all current imaged records to one standardized TIFF format.

b. Image declassified records using TIFF for archiving, TIFF 6 or PDF formats for access. Convert declassified versions of historically significant records to paper, microfilm or ASCII formats.

c. Image manuals, standards, directive type material using TIFF, ASCII and ASCII SGML or XML tagged files for archiving. Use PDF, HTML or XML formats for dissemination.

2. Plan and budget for migration of digital images every 3-5 years with of cost equivalent to 50 – 100% of the costs associated with original imaging project.

3. Convert documents that require long-term preservation from application format to an image format.

4. Develop standard set of access metadata of textual digital images using DoD 5015.2-STD, EAD, and Dublin Core as minimum set.

5. Work with Association for Information and Image Management (AIIM) and American National Standards Institute (ANSI) to standardize TIFF header data.

6. Work with NARA to:

a. Establish criteria for selection of digital images for accessioning in the National Archives

b. Accession digital images that have been imaged in the most stable format available and those that meet the selection criteria.

c. Establish guidelines describing metadata that must accompany digital image when submitted for archival accessioning

d. Study and evaluate migration strategies applied to digital data archives to application in the maintenance of textual digital images.

e. Study and evaluate formats designed for non-textual material, e.g. photography, aerial imagery, x-rays, radar, for compatibility with textual digital image formats in the archive environment.

f. Study and evaluate de jure interchange formats for long-term archive acceptance and application in the field.

7 References

The Applied Technologies Group, Enhancing Applications with Imaging Technology, The Technology Guide Series, 1998.

“Automatic Assessment of Document Image Quality,” The Communicator, Vol. III, No. 1, Summer 1998, http://dprc.seas.gwu.edu/dprc5/current_news/communicator/communicator_sum98.htm.

Behavioral Computational Neuropsychology (BCN) Group, Founding Committee, First General Announcement Proposing a Knowledge Processing Manhattan Project, May 24, 1998.

Carlin, John W., Statement on the Report of the Electronic Records Work Group Of the National Archives and Records Administration (NARA), 21 September 1998. http://www.nara.gov/records/grs20/state921.html.

Carlin, John W., “The Uncertain Future of the Past,” Washington Post, August 23, 1998, p. C3.

Carson, Steve, “Basic Image Interchange Format (BIIF),” GSC Associated Inc., http://www.acm.org/tsc/.

Cohen, Edmund; “Declassification and Records Management;” U.S. Central Intelligence Agency; Presentation, August 31, 1995.

Declassification Productivity Research Center, The George Washington University, Declassification Products, http://dprc.seas.gwu.edu/dprc5/declassification_products/products.htm.

Declassification Productivity Research Center, The George Washington University, Declassification Technologies, http://dprc.seas.gwu.edu/dprc5/declassification_tech.

Declassification Productivity Research Center, The George Washington University, Integrated Declassification for Executive Agencies (IDEA) System, http://dprc.seas.gwu.edu/dprc5/declassification_products/idea.htm.

DoD 5015.2-STD "Design Criteria Standard for Electronic Records Managements Software Applications," http://jitc.fhu.disa.mil/recmgt/rma-ps/index.html

EDIS, http://dprc.seas.gwu.edu/dprc5/declassification_products/EDIS_ver1.2.htm.

Electronic Imaging Standards for Archiving Records, report prepared by Logicon Communications Group for Deputy Assistant Secretary of Defense, Information Management, May 31, 1997.

Gray, Douglas E., "Preparing Graphics for The Web – Which Format to Use," May 25, 1997 http://www.servtech.com/~dougg/graphics/whichone.htm

Guidelines on the Management of Electronic Records from Office Systems. "Chapter 6: Transfer to the PRO," pg. 43-51, http://www.pro.gov.uk/recordsmanagement/eros/default.htm.

Hickok, Gene J.; "Media Life-Cycle Costs and Integrity;" National Media Laboratory; Presentation.

Information technology – Computer Graphics and image processing – Image Processing and Interchange (IPI) - Functional Specifition – Part 5: Basic Image Interchange Format (BIFF), ISMC – ISO?IEC 12087-5:1998(E), 1 December 1998. http://164.214.2.51/ntb/baseline/docs/biif/index.html

JPEG: http://src.doc.ic.ac.uk/media/visual/collections/funet-pics/jpeg/JPEG.FAQ.

Kenney, Anne R. and Rieger, Oya Y.; "Using Kodak Photo CD Technology for Preservation and Access," A Guide for Librarians, Archivists, and Curators; Department of Preservation and Conservation, Cornell University Library, May 1998.

National Library of Australia, "Preserving Access to Digital Information," http://www.nla.gov.au/padi/

"Preserving Digital Information, Final Report and Recommendations" by the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group, May 20, 1996, http://www.rlg.org/ArchTF/.

Prueitt, Paul S.; "Induction, Deduction and a New Form of Machine Intelligence;" Declassification Productivity Research Center, The George Washington University; April 8, 1998.

Prueitt, Paul S.; "Measurement, Categorization and Bi-level Computational Memory;" Declassification Productivity Research Center, The George Washington University; June 1, 1998.

Puglia, Steven and Roginski, Barry; "NARA Guidelines for Digitizing Archival Materials for Electronic Access;" National Archives and Records Administration; January 1998.

Raster Graphic Interchange Standards http://www2.echo.lu/oii/en/raster.html.

Research Library Group, "The RLG Worksheet for Estimating Digital Reformatting Costs," http://www.rlg.org/preserv/RLGtools.html

"A Strategic Policy Framework for Creating and Preserving Digital Collections," http://ahds.ac.uk/manage/framework.htm

"Technical Recommendations for Digital Imaging Projects," http://www.columbia.edu/acis/dl/imagespec.html

The VRML Repository, http://www.sdsc.edu/vrml

"Why all the argument about file formats?" http://www.faqs.org/faqs/jpeg-faq/part1/section-14.html.